SIGNAL AND NOISE IN GENE LENGTH DISTRIBUTIONS OF DECODED GENOMES

Aydin Tözeren[1], Neil D. Weston[1,2], and Zihang Ou[1]
[1]Department of Biomedical Engineering, The Catholic University of America, Washington DC 20064, USA; [2]Geosciences Research Laboratory; Silver Spring, MD 20910, USA

Abstract-**Development of objective measures for comparing genome data has become an important goal in computational genomics. Gene length distribution histograms derived from the predicted protein-coding gene populations of decoded genomes were considered as tools for genomic comparison. The first and second moments of the histograms map the decoded genomes as isolated points onto an Euclidean plane. The clustering of the genomic points on the plane preserves the separation of the three domains of life. The mapping of genomes onto a plane allows for the use of vector analysis in the study of genome evolution.**
Keywords-computationa genomics; informatics

I. INTRODUCTION

The number of decoded genomes from all domains of life is rapidly increasing. The value of such genome data for evolutionary studies lies in comparison. Previously, macromolecules involved in the transcription/translation apparatus were compared across genomes in the investigation of the evolution of species [1]. The life forms on Earth were grouped into three domains (bacteria, archaea, and eukarya) according to the similarity in the sequences of small rRNA subunits. The resulting phylogenetic tree separates bacteria from the lineage that later diverged into archaea and eukarya at its earliest branching. Recent studies tracing the time lines of different genes (proteins) brought results that are in conflict with this widely endorsed universal tree of life. Moreover gene histories do not trace the significant lateral transfer of genes between organisms. Such transfers were shown to occur between organisms that belong to different domains or subdomains [2].

In this study we have investigated the potential benefits of arranging predicted protein-coding genes of a genome in the order of increasing gene length. The most immediately obvious benefit is that the data then gains a geometric shape and thus become amenable to standard signal analysis. Moreover, using wavelet decomposition it may be possible to identify the impact on the signal of such factors as errors in gene annotation, lateral horizontal transfer, and adaptive gene loss. The method proposed in this article provides a powerful tool for determining the direction of evolutionary movement of species due to the forces of lateral gene transfer and adaptive gene loss.

II. METHODOLOGY

Sets of annotated proteins for the decoded genomes that were discussed in this article were retrieved from the site ftp://ncbi.nlm.nih.gov/genbank/genomes. The decoded genomes used in the study are composed of 6 archaea, 18 bacteria and 1 eukaryote. Modified lists for predicted protein-coding genes for A. pernix were obtained from Natale et al. (12) at ftp://ncbi.nlm.nih.gov/pub/koonin/Apernix. We used the MATLAB wavelet decomposition software to express gene length histograms in terms of a low-frequency approximation to the mean trend and 4 detail oscillatory signals, each having a different time scale [21]. We have used three different types of mother wavelets (Daubechies, Haar, and Shannon wavelets) in decomposition for each histogram, and the results were identical. The wavelet analysis used here decomposes the signal into a fourth-level approximation (A4) and four detail signals corresponding to different characteristic scales.

Frequency distributions of random variables can be represented by a set of distribution moments. In this study we considered the moments defined as follows:

$$M1 = \{Ó\,[n\,(L)\,L]\}/N \qquad (1)$$

$$M2 = \{\{Ó\,[n\,(L)\,L^2]\}/N\}^{0.5} \qquad (2)$$

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Signal and Noise in Gene Length Distributions of Decoded Genomes	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Biomedical Engineering The Catholic University of America Washington, DC 20064	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

In these equations, N denotes the number of protein coding genes in a genome, L is the gene length and n is the number of predicted protein-coding genes at that gene length. The summation is over all possible gene lengths ranging from 1 codon to N codons.

III. RESULTS AND DISCUSSION

Gene Length Histograms

Our study focused on the predicted protein-coding gene populations of six archaeal and eighteen bacterial genomes presented by the National Center for Biotechnology Information prior to January 15, 2001. Yeast was the only eukaryote considered in the study. Figure 1 shows gene length histogram of Aeropyrum pernix (ap2) predicted by Natale et al. [2].

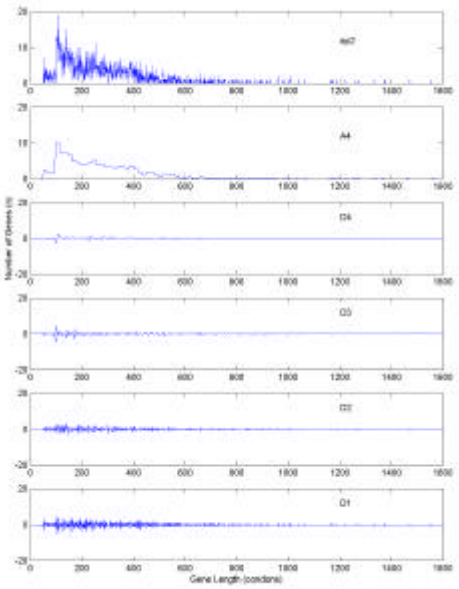

Figure 1 Length distribution of predicted protein-coding genes for Aeropyrum pernix (ap2) and its wavelet decomposition into a mean approximation (A4) and oscillatory detail components (D1 to D4).

Number of genes per gene length undergoes significant oscillations as the gene length increases one codon at a time from a minimum of about 50 codons to about 1400 codons. The oscillations depicted in D2 to D4 correspond to characteristic length scales increasing 2, 4, and 8 folds. The mean trend for Aeropyrum pernix is such that the number of genes decayed exponentially with increasing gene length following its peak around 101 codons. Note that the spike at about 101 codons is reflected in the mean approximation (A4) as well as in the detail curve (D1) with the lowest length scale. The figure also underscores the decreasing trend of the amplitude of oscillations with increasing characteristic length.

The histograms for decoded genomes could be partitioned into subsets according to geometric shape but in detail all were distinct from each other, and thus had the potential to be genomic signatures. The histogram for Escherichia coli shows no significant jump in the number of predicted proteins at 101 amino acids and the mean trend of the histogram resembles a skewed Gaussian distribution. Yeast, on the other hand, shows a significant jump at 101 codons but its genes are distributed over a much larger range of gene length than those that belong to Aeropyrum pernix. The discontinuity at 101 codons is quite apparent in some archaea including *Pyrococcus horikoshii* (phor). Its close phylogenic relative *Pyrococcus abysii* (paby) does not exhibit a sharp increase in the number of genes at this gene length. The biology behind the jump at 101 codons is not yet well understood.

Mapping of decoded genomes onto M1-M2 plane

We have determined the distribution moments M1 and M2 of the gene length distributions of decoded prokaryotes and yeast, as described in the methods section. Important features of a probability distribution can often be captured by a truncated set of distribution moments. The parameter M1 corresponds to the average gene length in the predicted protein-coding genes of a genome. The symbol M2 is the radius of gyration of the gene length histogram.

Figure 3 shows M1 plotted as a function of M2 for the predicted protein-coding gene populations of the decoded genomes considered in this study. Each decoded genome is represented as a point in the M1-M2 plane. Projection of predicted gene histograms onto a plane is reminiscent of the method in proteomics in which all known proteins are mapped as points on a plane using electrophoresis. The organisms to which the genomic points belong to are indicated in Fig. 2 by using the shorthand representations of the genome names, as presented in the methods section. The figure shows the domains of archaea and bacteria occupying different regions of the

M1-M2 plane. Yeast is positioned outside the regions occupied by archaea and bacteria.

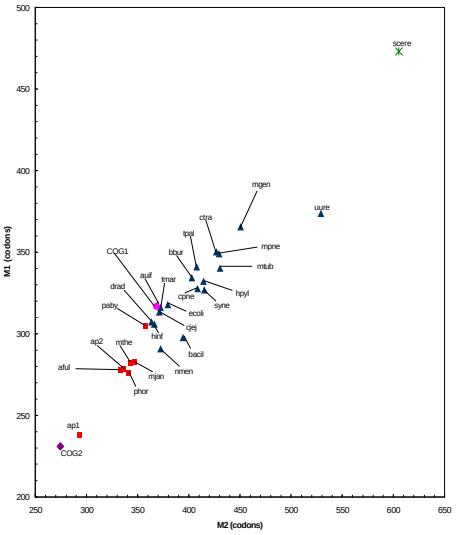

Figure 2 The spatial positions of decoded genomes on the M1-M2 plane. The coordinates of each genomic point were obtained by computing the mean gene length (M1) and length of gyration (M2) from the predicted gene length histogram of the decoded genomes. Species belonging to archaea are represented by red rectangles, bacteria by blue trangles, and yeast by a green star. The COG1 point is a purple star and and COG2 a purple circle.

The gene length histogram of the only completely decoded crenarchaeote, Aeropyrum pernix (ap1) is separated by the cluster of the 5 archaeal genomes from euryarchaeotes, which are on the other side of the deepest division of the archaea. However, when the alternative gene annotation predicted by Natale et al. [2] is used, this organism occupies a position on the M1-M2 plane (ap2) neighboring euryarchaeotes. Note also that Pyrococcus abysii is separated from this cluster and positioned closer to the interface between bacteria and archaea. The distance between Pyrococcus abysii and the rest of the decoded archaeal genomes may be due to the fact that Pyrococcus abysii possesses a larger set of metabolic genes than its phylogenic relative Pyrococcus horikoshii [12]. In its adaptation to heterotrophic life style, as a result of evolutionary forces, Pyrococcus horikoshii might have lost some of its metabolic pathways including that of aromatic amino acid pathway.

The genome coordinates shown in Fig. 2 reflect the present states of the organisms under consideration. The present positions of species on the M1-M2 plane are determined not only by phylogeny but also by the external evolutionary forces acting on them.

A point of reference to the distribution of decoded genomes on the M1-M2 plane could be obtained by computing the M1 and M2 values of the gene length histograms derived from the presently available COGs. We determined the minimum gene length and the average gene length for each COG. Then we created two histograms by determining the number of COGs that either have the same minimum gene length (COG1) or the same average gene length (COG2). Since each COG represents a cluster of ortholog genes, the resulting histograms represent different approximations to the actual gene length distribution of the so-called ancestor cells or cell clusters. Note that the histogram for COG1 is similar in shape to the histogram corresponding to Aeoropyrum pernix [3] whereas the histogram for COG2 exhibits a wider distribution around the mean, as is the case with Escherichia coli. Notice that the COG2 point falls in between the regions of archaea and bacteria in the M1-M2 plane whereas the COG1 is positioned further away in the domain of archaea (Fig. 2). If COG2 were to approximate the location of the ancestor cells on the M1-M2 plane, one would conclude that archaea might have moved toward lower left by primarily shedding genes as a result of evolutionary adaptation. Perhaps, the lateral acquisition of genes from archaea as well as bacteria was not enough to compensate for the drastic loss of genes from the ancestor gene pool. On the other hand, bacteria and eukaryotes might have increased their gene contents by gene duplication, lateral gene transfer, and other means. This raises the question whether COG1 or COG2 is a closer representation of the ancestor cell clusters. The answer must depend on the history of evolution of proteins. Is the shortest protein in a COG the most ancient? Recent genomic studies uncovered lineage specific expansions of domains and architectures of certain protein types in eukaryotes [4, 5]. The length of a protein carrying out a certain function would be expected to increase with the complexity of an organism due to increasing levels of regulation in protein expression.

IV. CONCLUSIONS

Collection of estimated protein-coding gene populations for each decoded genome comes to life and become visual objects when the population is presented as a distribution of the number of predicted genes as a function of gene length. These histograms comprise noisy signals that potentially contain important data concerning the evolution of species. Embedded in predicted gene length histograms is the genomic signature of an organism. The decomposition of the histogram signals provides a powerful technique for detecting errors in gene annotation and potentially identifying the laterally transferred gene clusters. The low frequency components of the histograms may be used to classify decoded genomes and shed light into their evolution. Protein-coding gene population distributions, when analyzed as nonlinear, stochastic signals, will likely to reveal further insights into the correlation between genomic structure, time, and environment.

When the average gene length M1 of the predicted protein-coding genes is plotted against the length of gyration parameter M2, the resulting dot plot separates the decoded genomes into clusters that reflect the three domains of life. The plot of genomic projections onto a plane allows for the computation of distances and angles between pairs of genomes. As the number of decoded genomes increases, M1-M2 plots of decoded genomes will capture the snap shot portrayal of the present-day organisms. Mapping of genomes onto a plane provides a visual aide as to spatial location of a newly decoded organism in relation to the organisms in various domains of life.

REFERENCES

[1] Doolittle WF: "Phylogenic classification and the universal tree." *Science*1999, **285**: 2124-2129.

[2] Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV: "Toward understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs)." *Genome Biology* 2000, **1**: 1-19.

[3]. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, et al: Complete genome sequence of an aerobic hyperthermohilic crenarchaeon Aeropyrum pernix K1. *DNA Res,* 1999, **6**, 83 -101.

[4]. Lander ES et al.: Initial sequencing and analysis of the human genome. *Nature*, 2001, **409**: 860-921.

[5]. Venter, JC et al. The Sequence of the Human Genome. *Science*, 2001, **291**: 1304-1351.